様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１９　日現在

機関番号：　３４４０６
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2016
課題番号：　２５３３０１４０

研究課題名（和文）音楽の自動擬音語変換を用いたクラシック音楽用検索システムの開発

研究課題名（英文）Development of music information retrieval system for instrumental music by using automatic music-to-onomatopoeias converter

研究代表者

鈴木　基之（Suzuki, Motoyuki）

大阪工業大学・情報科学部・准教授

研究者番号：３０２８２０１５

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究では擬音語による歌唱音声から楽曲を検索するシステムの開発を行った。音楽とその擬音語表現の関係を明らかにすることで，メロディ情報に加えて使用された擬音語表現の情報も用いて楽曲を検索する。こうしたシステムを実現するため，楽曲の擬音語表現への自動変換法，歌唱音声からの高精度なメロディ情報の抽出法と歌詞情報の高精度認識法，高速な楽曲検索法のそれぞれについて開発を行った。楽曲の擬音語変換は，音声認識の枠組みを応用することで妥当な表現への変換を可能とした。メロディ情報の抽出法と歌詞認識法は，それぞれ従来の方法と比較して高精度な認識を実現し，また高速検索法においては10%程度の計算量を削減した。

研究成果の概要（英文）：In this study, we have developed a music information retrieval (MIR) system for instrumental music, which uses singing voice by onomatopoeias as query.
Relationship between music and its representation by onomatopoeias can be used for retrieval in addition to melody information. In order to realize such MIR system, following four methods has been developed; automatic music-to-onomatopoeias converter, high performance melody information extractor and lyrics recognizer from singing voice, and high speed retriever.
From experimental results, the converter outputted appropriate onomatopoeias representations. The extractor and the recognizer outperformed conventional methods, and the retriever suppressed about 10% amount of calculation.

研究分野：　音声情報処理

キーワード：　楽曲検索システム　擬音語歌唱　音程抽出　歌唱音声認識

１．研究開始当初の背景

近年 Web を経由した音楽作品の購入が一般的となり，大量の音楽データから望みの曲を検索する機会も増大した。こうした場合，そのほとんどは曲名や歌手名といったメタ情報をキーとして検索が行われている。しかし，クラシック音楽に代表されるような歌唱音声の入っていない音楽（いわゆるインストゥルメンタル)は BGM 等で用いられることが非常に多く，「曲は知っているけど，曲名も作曲者も知らない」といったことが頻繁に起こる。こうした曲に対しては，従来の検索システムでは検索不可能である。

このような問題点に対し，研究代表者らを含むいくつかの研究グループでは，ハミングによる楽曲検索システムの開発を行ってきた。しかし音の高さや長さといった情報抽出法が未熟なために検索精度は決して高いものではなく，また歌唱方法も「タ・タ・タ」に限定されるなど，インターフェイスとしても自然なものではなかった。

２．研究の目的

一般にクラシック音楽等を歌唱する際には，「じゃじゃじゃじゃーん」といった擬音語による表現がよく用いられる。こうした擬音語表現は実に多彩である。例えば同じ音を演奏した場合でも，ピアノだと「ポン」，トランペットだと「プー」，鉄琴なら「コン」といったように，音色によって表現が異なる。また同じ楽器であっても，低い音なら「ポン」だが，高い音だと「ピン」，大きい音なら「ドン」でも小さい音だと「トン」など，音の高さや長さ，大きさなど，様々な要因によって表現が変化する。

こうした音と擬音語の関係について解明されれば，擬音語表現を検索キーとして用いることで音楽検索の精度を大きく向上させることが可能となる。しかし，音と擬音語の関係については純音など非常に単純なものに対してのみ調べられているに留まっている。そこで本研究では，楽曲データを擬音語表現に自動で変換する方法を提案し，擬音語歌唱による検索システムを実現する。更に音の高さや長さの抽出精度も向上させ，数千曲のデータベースに対して，実用に耐えうる精度での検索を実現する。

３．研究の方法

開発するシステムの全体像を図 1 に示す。本システムでは，元となる音楽データから，音符の高さや長さの系列情報（メロディ情報）のデータベースと， 擬音語表現へと変換したデータベースの 2 つを事前に準備しておく。

擬音語による歌唱音声が入力されると，そこからメロディ情報を抽出する。ここでは以前研究代表者らが開発した音程（2 つの音における高さの差）を高精度に抽出する方法を応用し，高精度なメロディ情報抽出法を新たに開発することで精度向上を目指す。

同時に，擬音語歌唱を音声認識することで，擬音語の系列を抽出する。この時，一般に歌唱音声は通常発声と異なることから，歌唱音声専用の音声認識器を新たに開発する。特に歌唱音声では通常音声と比較して発話長（音符の長さに対応）が大きく異なり，それが誤認識の主な原因となっていることから，先に抽出したメロディ情報と組み合わせることで認識精度の向上を計る。

最終的に，メロディ情報，擬音語系列双方の類似度を用いて，候補となる曲を検索する。

こうしたシステムを実現させるため，以下の項目について研究を行う。

(1) 楽曲の擬音語表現への変換法を確立する。擬音語と音との関係については，純音など単純な音についてのみ研究例が存在する。本研究においては，こうした要因別に関係を調べるのではなく，元の音楽データを音声波形，擬音語表現されたデータを発話内容の書き起こし文，と対応づけることで，通常の音声認識の枠組みをそのまま用いて統計的にモデル化を行う。こうすることで，すべての要因を内包した自動変換法を確立することが可能となる。

(2) 高精度なメロディ情報抽出法を開発する。音の高さはピッチ抽出に基づく方法を用いるのが一般的であるが，本研究では以前研究代表者らが開発したふたつの音符間の音程を直接推定する方法を応用し，最小二乗法を用いてより頑健に音程系列を推定する方法を開発する。

(3) 高精度歌唱音声認識法を開発する。具体的には，過去に開発した歌詞認識法と同様に歌唱特有の発話様式に音響モデルを適応させた上で，極端な発話長の違いによる挿入，脱落誤りに対処するため，音符の区切り時刻情報を用いた制限付き認識アルゴリズムを開発する。

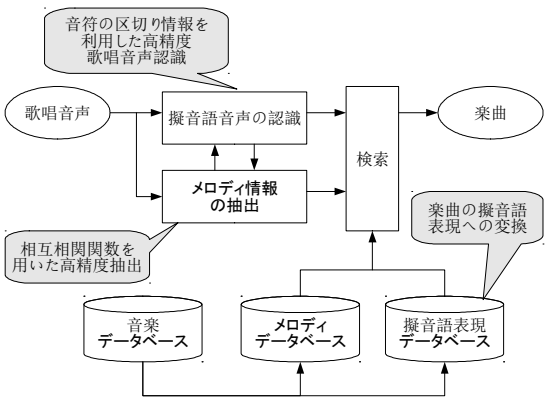

図 1 開発するシステムの全体像

(4) これまで開発してきた技術を統合し，擬音語歌唱入力による音楽検索システムを開発する。

４．研究成果

各研究項目について得られた研究成果は以下のとおりである。

(1) 楽曲の擬音語自動変換法の開発

楽曲を自動で擬音語表現に変換する方法を開発した。擬音語と音楽がどのような関係にあるのかについては，簡単な音での研究報告しかないため，まずは複雑な音楽と擬音語との関係を調査した。

まず，人間による擬音語歌唱データベースの整備を行った。比較的有名でなじみがあると思われるクラシック曲 20 曲を選定し，それぞれ代表的な部分（10 秒～20 秒程度）を 4 ヶ所ずつ切り出した。こうして得られた音楽データ 80 個を 49 名の協力者に聞かせ，それぞれ擬音語で歌唱してもらった。

こうして得られた擬音語歌唱データをテキストに書き起こし，どのような擬音語が用いられているか調査した。そもそも擬音語表現には最小単位となる「単語」が定義されていないため，階層 Pitman-Yor 言語モデルを用いた教師なし単語分割法を用いて頻出する擬音語系列を「単語」として抽出し，その使用傾向等を調査した。

「単語」は全部で 492 種類得られ，そのほとんどは 4 音節以内の長さであった。これらを使用頻度の観点から分析した（図 2）ところ，主要な楽器の種類によっておおまかに使用傾向が異なることがわかった。特にトランペットやフルート，ピアノといった楽器は使用される擬音語に偏りがあった。一方でバイオリンは様々な擬音語が使用されており，単純な関係ではない事がわかった。また，歌唱

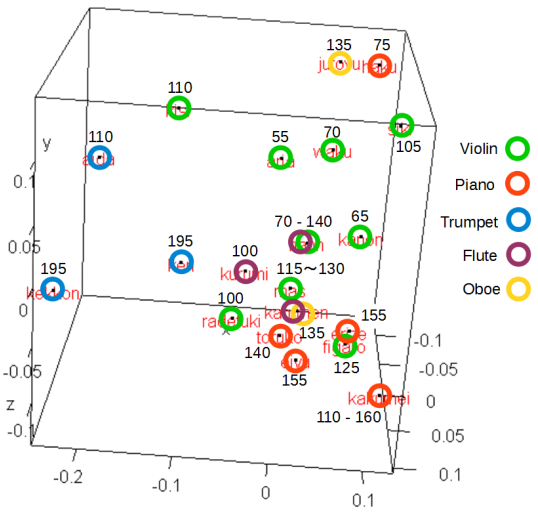

図 2 主要楽器ごとの擬音語使用傾向

者による擬音語表現の違いについては，様々な擬音語を使用する人やほとんど同じ擬音語しか利用しない人など，4 つのタイプに分類されることがわかった。

こうして得られた知見を元に，擬音語への自動変換法を開発した。自動変換法の基本的な枠組みは一般的な音声認識アルゴリズムと同じであり，楽曲を入力とし，擬音語テキストを出力とする。入力された楽曲はフレーム分割後，特徴量ベクトル系列へと変換される。こうして得られた特徴量ベクトル系列と事前に学習した HMM が比較され，擬音語テキストから得られた言語モデルとあわせて認識結果が出力される。

使用する特徴量は，音色に加えて音量や音高と擬音語との関係もあることから，通常音声認識で用いられる MFCC に加え，感情認識でよく用いられるパワー，ピッチ，Harmonic-to-Noise ratio，ゼロクロス比，Voice Quality といった値を用いた。

こうして変換された擬音語の妥当性を検証するための実験を行った。擬音語歌唱データベースを作成するのに用いた楽曲 80 データを自動で擬音語に変換し，その結果を人間による擬音語歌唱結果と比較した。その結果，極一部のみ同じ擬音語が使われていたが，ほとんどの部分は異なる擬音語へと変換されていた。

しかし，変換された擬音語を見ると，例えば人間による擬音語が「らー」であった場所が「たー」となっているなど，それなりに妥当であると思われる擬音語への変換が多く見られた。そこで，自動変換の結果得られた擬音語を元の楽曲と同時に示し，擬音語としての妥当性を人間に評価してもらう実験を行った。

自動変換の結果得られた擬音語テキストを画面上に表示し，元の楽曲の演奏にあわせてカラオケの歌詞表示のようにテキストの色を順次変化させるシステムを作成した。このシステムを用いて人間による評価を行ったところ，1 曲の半分以上の部分で妥当である，と判断された曲が 80 曲中 20 曲，部分的に妥当であると判断された曲まで含めると 53 曲が妥当な変換であると判断された。特に，ひとつの音符が長く演奏される部分は長音記号「—」に，またスタッカートのような奏法の部分は促音「っ」に，力強く演奏された音は濁音に変換されるなど，特徴的な演奏部分の多くは妥当な擬音語に変換されていることがわかった。

(2) 高精度メロディ情報抽出法の開発

歌唱音声中から高精度にメロディ情報，特に音高を抽出する方法の開発を行った。

音の高さはいわゆるピッチ抽出によって得ることができるが，時々倍音の周波数に間違えてしまうなど，その精度は必ずしも高くはない。そこで以前研究代表者らが開発したふたつの音符間の音程を直接推定する方法

を応用し，最小二乗法を用いてより頑健に音程系列を推定する方法を開発した。

あるふたつの音について，それぞれスペクトルを求め，その後周波数軸を対数に変換する。こうすることで，音の高さの違うふたつのスペクトルは，対数周波数軸方向への平行移動となる。そこで，ふたつのスペクトル間の対数周波数軸方向へのずれを計測すれば，音程を推定することが可能となる。この方法はスペクトル形状の類似性を元に音程を推定するため，特に調波構造のよく表れているスペクトル（主に有声音の区間）で有効であると思われる。この方法を応用し，音高系列を自動抽出する。

まず，入力された歌唱音声をフレームに分割し，それぞれ対数周波数軸に変換されたスペクトルを求める。次に，休符や息継ぎにあたると思われる無音に近いフレームと，無声子音にあたると思われる調波構造の少ないフレームを除外する。具体的には，無音部分の推定は平均パワーによる分類に加え，事前に無音部分，有音部分からそれぞれ学習していた GMM による尤度比も利用する。また調波構造の表れない区間については，Harmonic-to-Noise ratio の値をもとに推定する。

その後，残された（有声音にあたると思われる）すべてのフレームについて，あるフレームを基準として，その他のすべてのフレームとの間の音程を計算する，ということを，すべてのフレームを基準として繰り返す。こうして得られた音程系列は，本来すべて一致するはずであるが，実際にはノイズ等によって値が異なる。そこでこれらの系列から最小二乗法を用いて最もよい音程系列を推定する。

195 名による歌唱データベースを用いて音程の抽出実験を行ったところ，1 フレームあたりの平均推定誤差が 50.1cent と，半音の更に半分程度に抑えられていることがわかり，楽曲検索システム向けとしては十分な精度が得られていることがわかった。一方，従来からよく用いられているピッチ抽出に基づく方法では，精度がよいといわれている TANDEM-STRAIGHT を用いても平均推定誤差が 73.1cent と，今回開発した方法の方がより高い精度で推定できることもわかった。

(3) 高精度歌唱音声認識法の開発

一般に歌唱音声から歌詞情報を高精度に抽出することは困難であることが知られている。その原因のひとつは，歌唱することによって発声が通常音声と異なる事，もうひとつはメロディによって通常ではあり得ないほど長く発声するモーラが出現する，ということである。前者に対しては，歌唱音声を用いて音響モデルを適応化させる方法が提案されており，一定の効果が報告されている。一方後者については，言語モデルの制約を強くする方法が提案されているが，本質的な解決方法とは言えない。

歌唱音声における各モーラの継続時間長は，メロディに制約される。各モーラはメロディを表す音符に対応付いているため，長い音符に対応したモーラは長く，短い音符に対応したモーラは短く発話されることとなる。そこで，音符の区切り時刻の情報を用い，明示的に各モーラ長を制限することで，挿入誤りを低減させた歌唱音声認識法を開発した。

通常の音声データから得られる特徴量ベクトルとはかけ離れた値を持つ特殊な特徴量ベクトルを定義し，入力された歌唱音声から計算された特徴量ベクトル系列中の，音符の区切り時刻（この時刻は別途推定しておく）に対応するフレームにひとつずつ挿入していく。こうすることで，特徴量ベクトル系列でどの位置に音符の区切りがあるか，判別可能としておく。

また，音声認識を行う音響モデルに，挿入したベクトルに対応した特殊な音素 HMM をひとつ追加し，すべてのモーラ間で，この HMM を使うように発音辞書を変更する。こうすることで，認識仮説中のモーラの時間長を，音符長と一致させることが可能となる。

この方法では，音符の区切り時刻以外では，モーラ間の遷移が起こらない。つまり，ひとつの音符が必ずひとつのモーラに対応することになる。しかし，楽曲によってはひとつの音符が複数のモーラに対応したり，複数の音符がひとつのモーラに対応したりすることも考えられる。そこで，挿入するベクトルが持つ特殊な値を通常の音声特徴量に近づけることで制約を緩め，こうした 1 対多の対応を含む楽曲であっても高精度に認識するようにした。

48 曲の童謡を 27 名が歌唱した歌唱データ（全 198 データ）を用いて歌詞の認識実験を行った。歌唱特有の発声に対応するため，26 名の歌唱データで新たに音響モデルを構築し，残りの 1 名の歌唱データを認識する，という事を評価用歌唱者を変更しながら 27 回繰り返し，その平均値で評価した。

まず，音符とモーラを必ず 1 対 1 に対応させる設定で認識実験を行ったところ，楽曲中

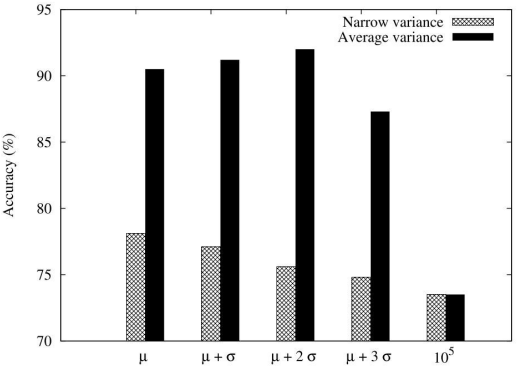

図 3 各種パラメータ設定と単語正解精度

に 1 対 1 の対応しかない曲については単語正解精度が 98.3%と，非常に高い精度を得ることができた。一方で 1 対多の対応を含む曲については認識精度が 40%台にまで落ちるなど，性能が大きく劣化した。198 データ全体での平均でみると，音符とモーラの対応関係に制限を加えていない従来からの方法での認識精度が 85.7%なのに対し，制限を加える方法では 73.5%と，精度の向上はできなかった。

そこで，音符とモーラの対応の制限を緩めた方法で認識を行った（図 3）。制限を緩めることで，1 対 1 の対応しかない曲に対しては認識精度が 97.5%となり，多少の性能劣化が見られた。一方で 1 対多対応を含む曲については大幅な性能の向上が見られ，最終的には，適切なパラメータを設定することで単語正解精度を 92.0%にまで向上させることができた。

(4) 楽曲検索システムの開発

高速高精度に楽曲を検索するシステムの開発を行った。楽曲検索システムにおいては，まず入力歌唱を特徴量系列へと変換し，データベースの特徴量と距離計算を行って検索する。一般には特徴量ベクトル間のユークリッド距離を用いることが多いが，この場合各次元間の類似関係を距離に反映させることができないため，入力歌唱の多少の変形による特徴量の変化に弱い。そこで，Earth Mover’s Distance（EMD）を距離尺度として用いる方法を採用した。

この方法では，入力歌唱を固定長のフレームに分割し，それぞれ特徴量を求める。それらとデータベース中の特徴量との間の距離計算を行うことで，最も類似した曲を検索している。しかしこの方法ではフレーム間の時間的な遷移は注目していないため，たまたま 1 フレームだけ類似した曲があると，それを検索結果としてしまう，という問題点があった。そこで，フレーム間の時間的遷移を制約条件として用い，検索精度を向上させた。

入力歌唱から得られたすべてのフレームと，データベース中の曲のすべてのフレーム間で EMD を計算し，その後フレームの時間的遷移を制約として，系列同士の距離を求める。こうして系列同士の距離が最も近い曲を検索結果とすることで，およそ 60%の精度で検索上位 10 位以内に正解の曲を挙げることができた。時間的制約を用いない場合はおよそ 30%の精度であったため，大幅に性能を向上させることができた。特に入力歌唱が長い時（フレーム系列が長い時）に時間的制約が有効に働くことがわかった。

また，EMD のもうひとつの問題点は，その計算時間の長さである。ユークリッド距離を計算するのと比較して，非常に計算負荷が高い。一般に楽曲検索システムでは，データベース中に何十万曲も登録されるため，高速に検索を行う技術の開発は非常に重要である。

そこで，データベース中に登録された特徴量を予めクラスタリングしておくことで代表的な特徴量集合を作り，それを用いて予備選択を行う方法を開発した。

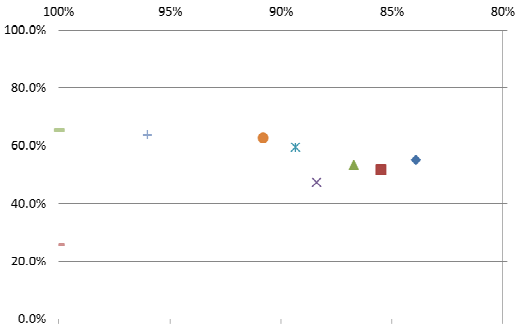

図 4 検索精度と計算量

まず，データベース中のすべての曲から得られた特徴量ベクトルをクラスタリングし，代表特徴量ベクトルをいくつか求めておく。これを用いてデータベース中の特徴量ベクトルをベクトル量子化することで，すべての曲は代表特徴量ベクトルだけを用いて表現することが可能となる。

こうして作成されたデータベースと入力歌唱をマッチングさせる。入力歌唱から得られた特徴量ベクトルはデータベース中のすべての特徴量との間で EMD を計算する必要があるが，データベース中の特徴量はすべて代表特徴量ベクトルに量子化されているため，ひとつの特徴量ベクトルは，代表特徴量ベクトルの数だけ EMD を計算すればよい。つまり，代表特徴量ベクトルの数を少なく設定しておけば，それだけ高速に距離計算を行うことが可能となる。

当然，代表特徴量ベクトルの数を減らせば，それだけ量子化誤差が増え，検索失敗へと繋がる。そこで，代表特徴量ベクトルを用いた計算結果で上位 n 曲を予備選択し，その後ベクトル量子化していない元々の特徴量を用いて距離を再計算する。こうすることで，高速化と高精度化を両立させた。

63 曲が登録されたデータベースを用い，27 名が童謡を歌唱したデータを検索した結果を図 4 に示す。ここで縦軸は検索結果上位 10 曲までに正解曲がはいっていた割合，横軸は予備選択を行わなかった時の計算量に対する割合を表す。いくつかの設定で検索実験を行ったところ，検索精度をほとんど落とさない設定では 10%程度の計算量の削減を達成した。また 15%程度計算量を削減しても，検索精度は 10 ポイント程度の低下に抑えられることがわかった。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　1　件）

(1) Motoyuki Suzuki, Akimitsu Hisaoka.

“Development of Singing-by-Onomatopoeia corpus for Query-by-Singing Music Information Retrieval system”, International Journal of Advanced Intelligence, Vol.9, No.1, pp.63-75. (2017) (査読有)

〔学会発表〕(計 6 件)

(1) 鈴木 基之, 杉田 裕亮.「音符区切り情報を用いた高精度歌唱音声認識」, 情報処理学会研究報告音楽情報科学 (MUS), 2017 年 6 月 17 日, お茶の水大学 (東京都・文京区).
(2) Motoyuki Suzuki, Akimitsu Hisaoka. “Development of Singing-by-Onomatopoeia corpus for Query-by-Singing Music Information Retrieval system”(Best paper award), 11th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE’16), 2016 年 12 月 15 日,沖縄県市町村自治会館 (沖縄県・那覇市)
(3) Motoyuki Suzuki, Kohei Kawashima. “Automatic motion selection method for spoken dialog scenario editor”, 20th Annual Conference on Knowledge-Based and Intelligent Information & Engineering Systems, 2016 年 9 月 5 日, ヨーク (英国)
(4) Motoyuki Suzuki. “Lyrics recognition from singing voice dealing with insertion error” (invited paper), RIEC International Symposium on Ultra-Realistic Interactive Acoustic Communications 2016, 2016 年 5 月 21 日, 宮城蔵王ロイヤルホテル (宮城県・刈田郡蔵王町)
(5) 鈴木 基之, 久保 勇人.「フレーム間の音程に注目した歌唱音声からのメロディ抽出法」, 情報処理学会 音楽情報科学研究会 (MUS), 2015 年 5 月 24 日, 電気通信大学 (東京都・調布市)
(6) 黒川 優太郎,佐藤 友哉,鈴木 基之.「擬音語による楽曲表現の分析と自動変換システムの開発」, 電子情報通信学会音声研究会, 2014 年 2 月 28 日, 徳島大学 (徳島県・徳島市)

６．研究組織

(1)研究代表者

鈴木 基之 (SUZUKI, Motoyuki)
大阪工業大学・情報科学部・准教授
研究者番号：30282015

(4)研究協力者

川島 滉平 (KAWASHIMA, Kohei)
久保 勇人 (KUBO, Hayato)
黒川 優太郎 (KUROKAWA, Yutaro)
駒井 里紀 (KOMAI, Riki)
佐藤 友哉 (SATO, Tomoya)
杉田 裕亮 (SUGITA, Yusuke)
杉本 侑太 (SUGIMOTO, Yuta)
久岡 昭允 (HISAOKA, Akimitsu)